



Early Journal Content on JSTOR, Free to Anyone in the World

This article is one of nearly 500,000 scholarly works digitized and made freely available to everyone in the world by JSTOR.

Known as the Early Journal Content, this set of works include research articles, news, letters, and other writings published in more than 200 of the oldest leading academic journals. The works date from the mid-seventeenth to the early twentieth centuries.

We encourage people to read and share the Early Journal Content openly and to tell others that this resource exists. People may post this content online or redistribute in any way for non-commercial purposes.

Read more about Early Journal Content at <http://about.jstor.org/participate-jstor/individuals/early-journal-content>.

JSTOR is a digital library of academic journals, books, and primary source objects. JSTOR helps people discover, use, and build upon a wide range of content through a powerful research and teaching platform, and preserves this content for future generations. JSTOR is part of ITHAKA, a not-for-profit organization that also includes Ithaka S+R and Portico. For more information about JSTOR, please contact support@jstor.org.

THE MATHEMATICAL REPRESENTATION OF FREQUENCY DISTRIBUTIONS

BY HARRY C. CARVER, *University of Michigan*

-
- SECTION I. DISTRIBUTIONS OF GRADUATED VARIATES
 SECTION II. DISTRIBUTIONS OF INTEGRAL VARIATES
 SECTION III. DIFFERENCE EQUATION GRADUATION*
 SECTION IV. APPLICATION OF THE HYPERGEOMETRIC SERIES*
-

$$\begin{aligned} \text{(A)} \quad u_x &= {}_{pn} C_{r-x} \quad {}_{qu} C_x \\ \text{(B)} \quad u_x &= {}_{pn} H_{r-x} \quad {}_{qu} H_x \end{aligned}$$

As an introduction to this paper on mathematical statistics the writer cannot refrain from quoting the first paragraph from the introduction of an extremely sound treatise on non-mathematical statistics, Zizek's *Statistical Averages*, a translation by Persons.†

"In nearly all branches of statistical investigation averages or means are very often used for various purposes of the greatest significance. Averages are, therefore, undoubtedly to be reckoned among the most important aids in statistical method. Even Guerry, the founder of moral statistics, indicated the extraordinary importance of averages in the following definition of statistics: 'The science of statistics consists essentially in the methodological enumeration of variable elements, whose mean it determines.' Edgeworth defines statistics as 'the science of those means which are presented by social phenomena'; and similarly Bowley says, 'Statistics may rightly be called the science of averages.' The application of averages has, as is well known, given rise to controversies of various kinds, which fill a considerable part of statistical literature. Not infrequently the incorrect use of averages has also led to erroneous conclusions and to contradictions which have shaken our confidence in statistics. Averages, indeed, are only applicable under strictly defined conditions, and conclusions based on averages are likewise permissible only within well defined limits. It is the task of statistical science to investigate the application and use of averages from the general methodological standpoint, and to determine the part which averages should play in statistical method."

In astronomy and in other fields in which precise measurements are necessary, the average of a series of independent measurements of equal weight on an unknown quantity is assumed to give the best approximation to, *i. e.*, the most probable value of, the magnitude in question. But the computation of the mean, instead of permitting a final answer to the problem, only gives rise to other questions of a more involved character. Thus, we may require the "probable error"

* Sections III and IV of this paper will appear in the September issue of the *QUARTERLY PUBLICATIONS*.

† New York: Henry Holt and Company, 1913.

of the mean in order to estimate the extent to which we may rely on this average, or if this mean is to be combined with other means, it is necessary to know the approximate "weight" of each in order to insure a proper combination of observations.

In statistics it is likewise of the utmost importance that we be able to measure the dependability and the significance of averages. Thus, it may be quite necessary that we know that the average or normal weight of a male infant sixteen months old is 23.25 pounds, or that the average weight of a man who is thirty-seven years old and six feet tall is 179 pounds; but it is likewise important that we know to what extent these averages may be relied upon and be able to state to what extent a sixteen months old infant weighing 19 pounds or a thirty-seven year old male six feet in height and weighing 200 pounds, is abnormal.

Again, if from past experience we observe that certain abnormal conditions of rainfall and temperature seem to be associated with abnormal crops, we immediately feel that it would be desirable to know to just what extent these phenomena are on the average associated, and to be able to determine, if possible, a functional relation between cause and effect. Assuming that a mathematical representation of the law of cause and effect is obtainable, we must hesitate to place any confidence on any dependent prediction until we have investigated the dependability of the mathematical predictions.

It is quite clear that the dependability of a mean must be a function of both the number of observations used in computing the mean and the magnitudes of the variations about the mean. It follows, therefore, that a critical investigation of the distribution of magnitude-determinations about their mean is essential.

The fundamental problem of mathematical statistics is to develop an analytical expression that is capable of representing any distribution of magnitude-determinations, *i. e.*, *variates*. The Gaussian or Normal Curve of Error, Pearson's Generalized Probability Curves, and Charlier's Type A and Type B curves are all the results of investigations of this fundamental problem.

In this paper we shall first investigate the general problem of distribution from the point of view of a general mathematical representation rather than from the point of view of the theory of probability. The writer believes it possible to justify this deviation from the usual method of presentation.

SECTION I

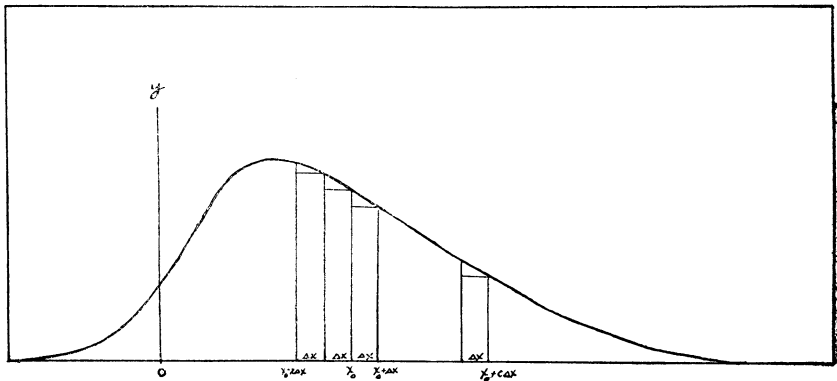
DISTRIBUTIONS OF GRADUATED VARIATES

Definition: *Graduated variates* are expressions of observed magnitudes which, from their nature, do not exist as integers but which may differ by infinitesimals.

We assume that variations from the mean follow some unknown law and that the observed class frequencies are approximations to corresponding areas under a theoretical *frequency function*, which affords a mathematical representation of the unknown law of distribution.

Let $f(x)$ be any frequency function which vanishes at $x = -l_1$ and $x = +l_2$, where the origin is taken at any point within the range. The limits may be either both finite, both infinite, or one finite and the other infinite.

Let the range $(-l_1, l_2)$ be divided into segments of equal length, Δx , and let ordinates be erected at each of the points of division. Now let rectangles be inscribed on these segments with altitudes equal to their respective right-hand boundaries.



Choosing one of these points of division as origin, the area of the rectangle whose right-hand boundary is at x_0 is

$$\Delta x \cdot f(x_0)$$

and that of any other rectangle whose right-hand boundary is $c \cdot \Delta x$ distant is

$$\Delta x \cdot f(x_0 + c \cdot \Delta x).$$

Let us now set up the function

$$\sum_{c=-\frac{l_1+x_0}{\Delta x}}^{\frac{l_2-x_0}{\Delta x}} \Delta x \cdot f(x_0+c \cdot \Delta x) e^{(x_0+c \cdot \Delta x)\omega i}$$

where $i = \sqrt{-1}$ and ω is a parameter, and observe that if we multiply each term of this summation by

$$e^{-x_0\omega i}$$

and integrate the resulting series term by term with respect to ω we obtain the following identity

$$\int_{-\pi/\Delta x}^{\pi/\Delta x} e^{-x_0\omega i} \left\{ \sum_{c=-\frac{l_1+x_0}{\Delta x}}^{\frac{l_2-x_0}{\Delta x}} \Delta x \cdot f(x_0+c \cdot \Delta x) e^{(x_0+c \cdot \Delta x)\omega i} \right\} d\omega = 2\pi f(x_0).$$

This follows immediately from the fact that the general term of the summation after multiplication by

$$e^{-x_0\omega i}$$

and subsequent integration yields, for $c \neq 0$

$$\Delta x \cdot f(x_0+c \cdot \Delta x) \int_{-\pi/\Delta x}^{\pi/\Delta x} (\cos c \cdot \Delta x \cdot \omega + i \sin c \cdot \Delta x \cdot \omega) d\omega = 0,$$

since c is necessarily an integer.

For $c \equiv 0$ we have

$$\Delta x \cdot f(x_0) \int_{-\pi/\Delta x}^{\pi/\Delta x} d\omega = 2\pi f(x_0).$$

We have therefore

$$(1) \quad f(x_0) = \frac{1}{2\pi} \int_{-\pi/\Delta x}^{\pi/\Delta x} e^{-x_0\omega i} \left\{ \sum_{c=-\frac{l_1+x_0}{\Delta x}}^{\frac{l_2-x_0}{\Delta x}} \Delta x \cdot f(x_0+c \cdot \Delta x) e^{(x_0+c \cdot \Delta x)\omega i} \right\} d\omega.$$

Now permitting Δx to approach zero we have

$$(2) \quad f(x_0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-x_0\omega i} \left\{ \int_{-l_1}^{l_2} f(x) e^{x\omega i} dx \right\} d\omega.$$

Formula (2) is merely a form of Fourier's Integral.

If we now place

$$(3) \quad \int_{-l_1}^{l_2} f(x) e^{x\omega i} dx = e^{c_0 + c_1 \omega i - c_2 \frac{\omega^2}{2}} \left\{ 1 + c_3 \frac{\omega^3 i^3}{\underline{3}} + c_4 \frac{\omega^4 i^4}{\underline{4}} + \dots \right\},$$

the coefficients, c_n , may be evaluated by expanding

$$\int_{-l_1}^{l_2} f(x) e^{-c_0 + \omega i(x - c_1) + c_2 \frac{\omega^2}{2}} dx$$

in a power series of ω and imposing the condition that the coefficients of ω^0 , ω^1 and ω^2 of the resulting series must equal 1, 0, and 0 respectively.

For values of c_n , ($n \geq 3$)

$$c_n i^n = e^{-c_0} D_{\omega}^{(n)} \int_{-l_1}^{l_2} f(x) e^{\omega i(x - c_1) + c_2 \frac{\omega^2}{2}} dx \Big|_{\omega=0},$$

and applying the operator formula

$$\phi(D) e^{ax} V = e^{ax} \phi(D + a) V$$

we may write

$$(4) \quad c_n i^n = e^{-c_0} \int_{-l_1}^{l_2} f(x) e^{\omega i(x - c_1)} \left\{ (D_{\omega} + i \overline{x - c_1})^n e^{c_2 \frac{\omega^2}{2}} \right\} dx \Big|_{\omega=0} \\ = i^n e^{-c_0} \int_{-l_1}^{l_2} \left\{ (x - c_1)^n - \frac{n^{(2)}}{2} c_2 (x - c_1)^{n-2} + \frac{n^{(4)}}{2^2} c_2^2 (x - c_1)^{n-4} \right. \\ \left. + \dots \right\} f(x) dx,$$

where

$$n^{(r)} = n(n-1)(n-2) \dots (n-r+1).$$

Imposing the condition that (4) must equal 1, 0, and 0 for $n=0$, 1, and 2 respectively, we have

$$e^{-c_0} = \frac{1}{\int_{-l_1}^{l_2} f(x) dx},$$

$$c_1 = \frac{\int_{-l_1}^{l_2} x f(x) dx}{\int_{-l_1}^{l_2} f(x) dx},$$

$$c_2 = \frac{\int_{-l_1}^{l_2} (x - c_1)^2 f(x) dx}{\int_{-l_1}^{l_2} f(x) dx}.$$

We see, then, that c_1 is the distance from the arbitrary origin to the mean and that c_2 is therefore the second moment of unit frequency about the mean, or in other words is the square of the standard deviation.

Representing, as is customary, by

b , the distance from the arbitrary origin to the mean,

σ , the standard deviation,

N , the total frequency of the distribution,

ν'_n , the n^{th} moment of unit frequency about the arbitrary origin,

$$\text{i.e. } \nu'_n = \frac{\int_{-l_1}^{l_2} x^n f(x) dx}{\int_{-l_1}^{l_2} f(x) dx},$$

ν_n , the n^{th} moment of unit frequency about the mean

$$\text{i.e. } \nu_n = \frac{\int_{-l_1}^{l_2} (x - c_1)^n f(x) dx}{\int_{-l_1}^{l_2} f(x) dx}$$

we obtain,

$$(5) \quad \nu_n = \nu'_n - {}_n C_1 b \nu'_{n-1} + {}_n C_2^2 b^2 \nu'_{n-2} - \dots$$

$$c_n = \nu_n - \frac{n^{(2)}}{2} \sigma^2 \nu_{n-2} + \frac{n^{(4)}}{2^2 [2]} \sigma^4 \nu_{n-4} - \frac{n^{(6)}}{2^3 [3]} \sigma^6 \nu_{n-6} + \dots$$

Referring to equations (2) and (3) we see that

$$f(x_0) = \frac{N}{2\pi} \int_{-\infty}^{\infty} e^{-(x_0-b)\omega i - \frac{\sigma^2 \omega^2}{2}} \left\{ 1 + c_3 \frac{\omega^3 i^3}{\underline{3}} + c_4 \frac{\omega^4 i^4}{\underline{4}} + \dots \right\} d\omega.$$

If we now set

$$\phi_{x_0} = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-(x_0-b)\omega i - \frac{\sigma^2 \omega^2}{2}} d\omega$$

and note that

$$\phi_{(x_0)}^{(n)} = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-(x_0-b)\omega i - \frac{\sigma^2 \omega^2}{2}} (-\omega i)^n d\omega$$

we have finally, dropping the subscript on x , that any ordinate

$$(6) \quad f(x) = N \left\{ \phi_x - \frac{c_3}{\underline{3}} \phi_x^{(3)} + \frac{c_4}{\underline{4}} \phi_x^{(4)} - \frac{c_5}{\underline{5}} \phi_x^{(5)} + \dots \right\},$$

where from equation (5)

$$(7) \quad \begin{cases} c_3 = \nu_3 \\ c_4 = \nu_4 - 3\sigma^4 \\ c_5 = \nu_5 - 10\sigma^2\nu_3 \\ c_6 = \nu_6 - 15\sigma^2\nu_4 + 30\sigma^6 \\ c_7 = \nu_7 - 21\sigma^2\nu_5 + 105\sigma^4\nu_3 \\ c_8 = \nu_8 - 28\sigma^2\nu_6 + 210\sigma^4\nu_4 - 315\sigma^8 \\ \text{etc.} \end{cases}$$

From the well-known definite integral

$$\int_{-\infty}^{\infty} e^{-a^2 x^2} \cos bx \, dx = \frac{\sqrt{\pi}}{a} e^{-\frac{b^2}{4a^2}}$$

we see that

$$(8) \quad \phi_x = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-b)^2}{2\sigma^2}}$$

Since the coefficients, c_n , are functions of the unknown law of distribution it would appear at first thought that we are now as far from having obtained a workable analytical expression for any distribution of graduate variates as we were at the start.

But although fluctuations in the values of individual class frequencies may arise as the result of "random sampling," we assume that the values of ν_n as computed from any observed frequency distribution

will approximately equal the corresponding values which are functions of the *unknown* law of distribution. Certain theorems involving the theory of "probable error" enable us to ascertain the degree of approximation involved in this assumption.

It follows, then, that if equation (6) is so rapidly convergent that terms after the third or fourth may be neglected, it affords us a simple representation of any distribution.

The rapidity with which this series converges, however, depends upon the extent to which the generating function, ϕ_x , is a fair approximation of the unknown law of distribution. In general, if the distribution be approximately symmetrical, equation (6) is very rapidly convergent. If, however, the distribution be bimodal or extremely asymmetrical, series (6) as a function of the symmetrical generating function, ϕ_x , is for practical purposes of no value because

(a) the number of required terms in (6) would necessitate an impracticable amount of labor; and

(b) the probable errors of c_n for values of n greater than 5 or 6 are generally so large that the assumption that, for these, we may substitute the moments computed from the observed data for the theoretical moments, is open to serious criticism.

It should be understood, however, that theoretically equation (6) is capable of representing any frequency function associated with graduated variates, but that convergence means one thing and rapid convergence quite another.

In case the distribution is skew we may proceed as follows:

Let θ_x be any arbitrarily chosen function which may be used as a statistical generating function and which, being a frequency function, may be expressed by means of equation (6) in the series

$$(9) \quad \theta_x = M \left\{ \phi_x - \frac{d_3}{\underline{3}} \phi_x^{(3)} + \frac{d_4}{\underline{4}} \phi_x^{(4)} - \frac{d_5}{\underline{5}} \phi_x^{(5)} + \dots \right\},$$

where d_n is a function of the moments of θ_x .

Eliminating ϕ_x between (6) and (9) we obtain

$$(10) \quad f(x) = \frac{N}{M} \left\{ \theta_x - \frac{A_3}{\underline{3}} \theta_x^{(3)} + \frac{A_4}{\underline{4}} \theta_x^{(4)} - \frac{A_5}{\underline{5}} \theta_x^{(5)} + \dots \right\}$$

where

$$\begin{aligned} A_3 &= c_3 - d_3 \\ A_4 &= c_4 - d_4 \\ A_5 &= c_5 - d_5 \\ A_6 &= (c_6 - d_6) - 20d_3(c_3 - d_3) \\ &\dots \end{aligned}$$

The area under $f(x)$ between any two limits may now be computed since

$$(11) \quad \int_a^b f(x)dx = \frac{N}{M} \left[\int \theta_x dx - \frac{A_3}{\underline{3}} \theta_x^{(2)} + \frac{A_4}{\underline{4}} \theta_x^{(3)} - \dots \right].$$

Inasmuch as tables for both the definite integral from 0 to x and $D^{(n)}$ ($n=1, 2, 3, 4, 5, 6$) have been computed for

$$\phi_x = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-b)^2}{2\sigma^2}}$$

it follows that in problems for which ϕ_x may be applicable, we are in a position to proceed with numerical computations.

It is highly desirable, however, that investigations be made in the endeavor to establish a reference list of generating functions possessing various desired properties. If the indefinite integral of any of these generating functions could readily be computed, such an investigation would be of the greatest value.

Again, since in the making of computations it would be much easier to obtain the numerical value of the n^{th} difference of θ_x than for the n^{th} derivative, it will be found of frequent advantage to utilize the following relation

$$D^n = \left(\Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} - \dots \right)^n$$

and write (10) and (11) as

$$(12) \quad f(x) = \frac{N}{M} \left\{ \theta_x - \frac{B_3}{\underline{3}} \Delta^3 \theta_x + \frac{B_4}{\underline{4}} \Delta^4 \theta_x - \dots \right\} \text{ and}$$

$$(13) \quad \int_a^b f(x)dx = \frac{N}{M} \left[\int \theta_x dx - \frac{B'_3}{\underline{3}} \Delta^2 \theta_x + \frac{B'_4}{\underline{4}} \Delta^3 \theta_x - \dots \right] \text{ respectively,}$$

where

$$B_3 = A_3$$

$$B'_3 = A_3$$

$$B_4 = A_4 + 6A_3$$

$$B'_4 = A_4 + 4A_3$$

$$B_5 = A_5 + 10A_4 + 35A_3$$

$$B'_5 = A_5 + \frac{15}{2}A_4 + \frac{55}{3}A_3$$

$$B_6 = A_6 + 15A_5 + 85A_4 + 225A_3$$

$$B'_6 = A_6 + 12A_5 + \frac{105}{2}A_4 + 100A_3$$

...

...

In order that formulas 12 and 13 may be useful, it is necessary that the class interval be taken as the unit of x for θ_x and for all computed moments.

SECTION II

DISTRIBUTIONS OF INTEGRAL VARIATES

Definition: *Integral variates* are expressions of observed magnitudes which, from their nature, exist only as integers. For example: number of petals of a flower, number of persons in a family, etc.

We assume that variations from the mean follow some unknown law and that the observed class frequencies are approximations to corresponding *ordinates* of the theoretical frequency function which affords a mathematical representation of the unknown law of distribution.

We may readily write down an expression from which we may develop the general law of distribution for integral variates by placing c equal to unity in formula (1) of the preceding section. We have, then,

$$(14) \quad f(x_0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-x_0 \omega i} \left\{ \sum_{x=-l_1}^{l_2} f(x) e^{x \omega i} \right\} d\omega.$$

Placing now

$$\sum_{x=-l_1}^{l_2} f(x) e^{x \omega i} = e^{c_0 + c_1 \omega i - c_2 \frac{\omega^2}{2}} \left\{ 1 + c_3 \frac{\omega^3 i^3}{3} + c_4 \frac{\omega^4 i^4}{4} + \dots \right\}$$

and proceeding exactly as before, we obtain, corresponding to (6),

$$(15) \quad f(x) = N \left\{ \phi_x - \frac{c_3}{3} \phi_x^{(3)} + \frac{c_4}{4} \phi_x^{(4)} - \dots \right\}$$

where now the values of c_n are the same as those given in (7) if we now define

$$\nu_n = \frac{\Sigma (x - c_1)^n f(x)}{\Sigma f(x)}$$

and

$$(16) \quad \phi_x = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-(x_0 - b) \omega i - \frac{\sigma^2 \omega^2}{2}} d\omega.$$

This ϕ_x may be eliminated as before, producing expansions corresponding to (10) and (12).

DISCUSSION OF SECTIONS I AND II

In *Über das Fehlergesetz* and *Die zweite Form des Fehlergesetzes*, Charlier has shown that on the hypothesis of elementary errors any law of error may assume one of the two forms.

Type A

$$\phi_x + A_3\phi_x^{(3)} + A_4\phi_x^{(4)} + \dots$$

Type B

$$\psi_{(x)} + B_1\Delta\psi_{(x)} + B_2\Delta^2\psi_{(x)} + \dots$$

where

$$\phi_x = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-(x_0-b)\omega i - \frac{\sigma^2\omega^2}{2}} d\omega,$$

and

$$\begin{aligned} \psi_{(x)} &= \frac{e^{-\lambda}}{\pi} \int_0^{-\pi} e^{x \cos \omega} \cos (\lambda \sin \omega - Z\omega) d\omega \\ &= \frac{e^{-\lambda} \lambda^x}{\lfloor x \rfloor} \end{aligned}$$

if x be an integer.

Owing to an oversight, π instead of ∞ appears as each limit of ϕ_x . The constants A_n and B_n are determined in *Über die Darstellung willkürlicher Funktionen* by applications of the method of moments, but appear to differ from the ones developed herein for values of n greater than (5) since by our general formula (5),

$$c_6 = \nu_6 - 15\sigma^2\nu_4 + 30\sigma^6,$$

whereas Charlier gives

$$c_6 = \nu_6 - 15\sigma^2\nu_4 + 15\sigma^6.$$

For values of n less than 6 the results agree.

As the result of the present investigation it would appear that there is no essential difference between Charlier's Type A and Type B curves, aside from the fact that both are merely special cases of formulas (10) and (12) when we define our generating functions as

$$\theta_x = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-b)^2}{2\sigma^2}}$$

and

$$\theta_x = \frac{e^{-\lambda} \lambda^x}{\lfloor x \rfloor} \text{ respectively.}$$

The fact that the second generating function may assume asymmetrical as well as nearly symmetrical forms renders the Type B curves more powerful than those of Type A.

It would appear, however, that for graduated variates the continuous limit of Poisson's Exponential Binomial Limit, namely, Pearson's Type III curve,

$$y = y_0 e^{-\gamma x} \left(1 + \frac{x}{a}\right)^{\gamma a}$$

placing $\nu_2 = \nu_3$, might be more logically applied.

Again, if we do not impose on the above generating function the condition that $\nu_2 = \nu_3$, it will be noted that any law of distribution may be written,

$$f(x) = N \left\{ \theta_x + \frac{A_4}{\underline{4}} \theta_x^{(4)} - \frac{A_5}{\underline{5}} \theta_x^{(5)} + \dots \right\},$$

where θ_x is the above mentioned Pearson's Type III function. Again we might use

$$f(x) = N \left\{ \theta_x + \frac{B_4}{\underline{4}} \Delta^4 \theta_x - \frac{B_5}{\underline{5}} \Delta^5 \theta_x + \dots \right\}.$$

Ordinarily

$$f(x) = N \left\{ \theta_x + \frac{B_4}{\underline{4}} \Delta^4 \theta_x \right\}$$

will give practically the same results as any one of Pearson's Generalized Probability Curves, and will involve far less labor.